Statistics and Probability Letters 144 (2019) 16–22

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Copula-based segmentation of cylindrical time series

Francesco Lagona

University of Roma Tre Via G. Chiabrera 199, 00145, Rome, Italy

ARTICLE INFO

Article history:
Available online 20 April 2018

MSC:
62P12
62M05

Keywords:
Cylindrical data
Copula
Hidden Markov model
Wrapped Cauchy
Weibull

ABSTRACT

A hidden Markov model is proposed for segmenting cylindrical time series according to a finite number of latent classes, associated with copula-based cylindrical densities. It provides a parsimonious and computationally tractable approach that integrates circular–linear correlation, multimodality and temporal auto-correlation.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Bivariate sequences of angles and intensities are often referred to as cylindrical time series, because the pair of an angle and an intensity can be represented as a point on a cylinder. In environmental studies, examples of these data include time series of wind directions and pollutant concentrations (García-Portugués et al., 2013), wind directions and speeds (Mastrantonio et al., 2015) and wave directions and heights (Lagona et al., 2015b).

The analysis of cylindrical time series is complicated by the difficulties in modeling the dependence between angular and linear measurements and the temporal correlation between cylindrical observations. An additional complication that often arises in environmental studies is the multi-modality of the marginal distribution of the data because environmental cylindrical data are typically observed under heterogeneous conditions that vary over time.

This paper introduces a parsimonious hidden Markov model (HMM) that simultaneously accounts for linear–circular dependence, temporal auto-correlation and multi-modality. Under this model, the distribution of cylindrical data is approximated by a mixture of copula-based cylindrical densities, whose parameters depend on the evolution of a latent Markov chain. While the copula-based cylindrical density accommodates linear–circular dependence, a mixture of copula-based densities allows for multimodality and, finally, a latent Markov chain accounts for temporal correlation.

This model extends previous proposals that are either based on mixtures of conditionally independent linear and circular densities (Holzmann et al., 2006) or based on mixtures of Abe–Ley cylindrical densities (Lagona et al., 2015a). It provides an intuitively appealing framework where the data are modeled by integrating conventional tools of statistical analysis: a copula, a mixture and a Markov chain. It is furthermore numerically tractable, by exploiting a suitable Expectation Maximization (EM) algorithm for parameter estimation.

The rest of the paper is organized as follows. Section 2 briefly describes the copula-based cylindrical distributions and introduces the proposed model. Section 3 is devoted to maximum likelihood parameter estimation and Section 4 illustrates the proposal on a case study of wave heights and directions. Relevant points of discussion are finally summarized in Section 5.

E-mail address: francesco.lagona@uniroma3.it.

https://doi.org/10.1016/j.spl.2018.04.011
0167-7152/© 2018 Elsevier B.V. All rights reserved.

2. A hidden Markov model with copula-based cylindrical densities

Copula-based cylindrical distributions. A cylindrical sample is a pair $\mathbf{z} = (x, y)$, where $x \in [0, 2\pi)$ is a point in the circle and y is a point in the positive semi-line $[0, +\infty)$. There are several ways to specify a parametric density on the cylinder (Ley and Verdebout, 2017). A simple approach relies on copulas (Jones et al., 2015): the joint distribution of two variables is specified as a function of the two univariate marginal distributions. More precisely, let $f(x; \alpha)$ be a density on the circle, known up to a parameter α, with cumulative distribution function (cdf) $F(x; \alpha)$, defined with respect to a fixed, although arbitrary, origin. Moreover, let $f(y; \beta)$ be a density on the semi-line, known up to a parameter β, with cdf $F(y; \beta)$. Finally, let $g(v; \gamma)$, $v \in [0, 2\pi)$ be a parametric circular density, known up to a parameter γ. Then, Johnson and Wehrly (1978) prove that

$$f_q(x, y; \theta) = 2\pi g\,[2\pi F(x; \alpha) - q2\pi F(y; \beta); \gamma]f(x; \alpha)f(y; \beta) \quad q = \pm 1 \tag{1}$$

is a parametric cylindrical density with support $[0, 2\pi) \times [0, +\infty)$, known up to the parameter vector $\theta = (\alpha, \beta, \gamma)$, having the marginal densities $f(x; \alpha)$ and $f(y; \beta)$. Eq. (1) is a typical example of a copula-based construction of a bivariate density, obtained by de-coupling the margins from the joint distribution, which is specified by means of the circular density g. When the binding density g is the uniform circular distribution, say $g(x) = (2\pi)^{-1}$, then Eq. (1) reduces to the product of the marginal densities. Otherwise, the dependence between x and y is captured by the concentration of g: when g is highly concentrated, the dependence is high; when g is more diffuse, dependence is low. Eq. (1) actually defines two densities, indexed by the value of q, which determines whether the dependence between x and y is positive ($q = 1$) or negative ($q = -1$).

A cylindrical hidden Markov model. Let $\mathbf{z} = (\mathbf{z}_t, t = 1, \ldots, T)$, $\mathbf{z}_t = (x_t, y_t)$, $x_t \in [0, 2\pi)$, $y_t \in [0, +\infty)$ be a cylindrical time series. We assume that the distribution of the data is driven by the evolution of an unobserved Markov chain with K states, which represents (time-varying) latent classes and can be specified as a sequence $\mathbf{u} = (\mathbf{u}_t, t = 1, \ldots, T)$ of multinomial variables $\mathbf{u}_t = (u_{t1} \ldots u_{tK})$ with one trial and K classes, whose binary components represent class membership at time t. The joint distribution $p(\mathbf{u}; \pi)$ of the chain is fully known up to a parameter π that includes K initial probabilities $\pi_k = P(u_{1k} = 1)$, $k = 1, \ldots, K$, $\sum_k \pi_k = 1$, and K^2 transition probabilities $\pi_{hk} = P(u_{tk} = 1 | u_{t-1,h} = 1)$, $h, k = 1, \ldots, K$, $\sum_k \pi_{hk} = 1$. Formally, we assume that

$$p(\mathbf{u}; \pi) = \prod_{k=1}^{K} \pi_k^{u_{1k}} \prod_{t=2}^{T} \prod_{h=1}^{K} \prod_{k=1}^{K} \pi_{hk}^{u_{t-1,h}u_{tk}}. \tag{2}$$

The specification of the HMM is completed by assuming that the observations are conditionally independent, given a realization of the Markov chain. As a result, the conditional distribution of the observed process, given the latent process, takes the form of a product density, say

$$f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K) = \prod_{t=1}^{T} \prod_{k=1}^{K} f_q(\mathbf{z}_t; \theta_k)^{u_{tk}}, \tag{3}$$

where $f(\mathbf{z}; \theta_k)$, $k = 1, \ldots, K$ are the K cylindrical densities defined by (1) and known up to a vector of parameters $\theta_k = (\alpha_k, \beta_k, \gamma_k)$.

The likelihood function of the model is therefore obtained by integrating the joint density of the observed data and the unobserved class memberships with respect to the segmentation $\mathbf{u}$, namely

$$L(\pi, \theta; \mathbf{z}) = \sum_{\mathbf{u}} p(\mathbf{u}; \pi) f(\mathbf{z}|\mathbf{u}; \theta_1, \ldots, \theta_K). \tag{4}$$

By computing the maximum likelihood estimate $\hat{\theta}$, the cylindrical time series can be then segmented according to the posterior probabilities of class membership

$$\hat{\pi}_{tk} = P(u_{tk} = 1 \mid \mathbf{z}; \hat{\theta}), \tag{5}$$

based on $\hat{\theta}$. More precisely, the observation at time t can be allocated to class $k^\star$ if $\hat{\pi}_{tk^\star} \geq \hat{\pi}_{th}$, for each $h = 1 \ldots K$ (maximum a posterior, MAP, allocation).

When the transition probability matrix has equal rows, the model reduces to a latent class model where observations are clustered by ignoring the information redundancy that is due to temporal correlation. More generally, the proposed HMM segments the series by accounting for similarities not only in the variable space but also in a temporal neighborhood.

3. Parameter estimation

An EM algorithm can be exploited to maximize the likelihood function (4). It is based on the following complete-data log-likelihood function

$$\log L_{\text{comp}}(\theta, \mathbf{u}, \mathbf{z}) = \sum_{k=1}^{K} u_{1k} \log \pi_k + \sum_{t=2}^{T} \sum_{h=1}^{K} \sum_{k=1}^{K} u_{t-1,h} u_{t,k} \log \pi_{hk} + \sum_{t=1}^{T} \sum_{k=1}^{K} u_{tk} \log f(\mathbf{z}_t; \theta_k). \tag{6}$$

The algorithm is iterated by alternating an expectation (E) and a maximization (M) step. Given the estimates $\hat{\pi}_s$ and $\hat{\theta}_s$, obtained at the end of the sth iteration, the $(s+1)$th iteration is initialized by the E-step, which evaluates the expected value of the complete data log-likelihood (6) with respect to the conditional distribution of the missing values u_{tk} given the observed data.

The E step reduces to the computation of the univariate posterior probabilities of each latent state at time t, $\hat{\pi}_{tk} = P(u_{tk} = 1 \mid \mathbf{z}, \hat{\pi}_s, \hat{\theta}_s) \quad k = 1 \dots K, t = 1 \dots T$, and the computation of the bivariate posterior probabilities of each pair of states in two adjacent times, say $\hat{\pi}_{t-1,t,hk} = P(u_{t-1,h} = 1, u_{tk} = 1 \mid \mathbf{z}, \hat{\pi}_s, \hat{\theta}_s) \quad h, k = 1 \dots K, t = 2 \dots T$. The task of computing these posterior probabilities from an estimate $(\hat{\pi}_s, \hat{\theta}_s)$ is generally referred to as the HMM-smoothing numerical issue and it is typically solved by specifying the posterior probabilities in terms of suitably normalized functions, which can be computed recursively, avoiding unpractical summations over the state space of latent Markov chain and numerical under- and over-flows. In this paper, we exploited the HMM-smoothing algorithm that is described by Bulla et al. (2012).

The M-step of the algorithm updates the estimate $(\hat{\pi}_s, \hat{\theta}_s)$ with a new estimate $(\hat{\pi}_{s+1}, \hat{\theta}_{s+1})$, by maximizing the expected value of the complete data log-likelihood (6), obtained from the previous E step. This expected value is the sum of functions that depend on independent sets of parameters and can therefore be maximized separately. Maximization with respect to the transition probabilities π_{hk}, under the constraints $\sum_{k=1}^{K} \pi_{hk} = 1$, $h = 1 \dots K$, provides the closed-form updating formula

$$\hat{\pi}_{hk(s+1)} = \frac{\sum_{t=2}^{T} \hat{\pi}_{t-1,t,hk}(\hat{\pi}_s, \hat{\theta}_s)}{\sum_{t=2}^{T} \hat{\pi}_{t-1,h}(\hat{\pi}_s, \hat{\theta}_s)}, \qquad h, k = 1, \dots, K.$$

Maximization with respect to the parameters θ_k of the kth copula-based cylindrical components reduces to maximize

$$\sum_{t=1}^{T} \hat{\pi}_{tk} f(\mathbf{z}_t; \theta_k). \tag{7}$$

We can maximize (7) with respect to all the parameters or, more efficiently, we can take an IFM (Kim et al., 2007, inference function for margins) approach. Precisely, (7) can be written as the sum of three components, namely

$$\sum_{t=1}^{T} \hat{\pi}_{tk} f(\mathbf{z}_t; \theta_k) = \sum_{t=1}^{T} \hat{\pi}_{tk} \log g(2\pi(F(x_t; \alpha) - qF(y_t; \beta)); \gamma) \tag{8}$$

$$+ \sum_{t=1}^{T} \hat{\pi}_{tk} f(x_t; \alpha_k) \tag{9}$$

$$+ \sum_{t=1}^{T} \hat{\pi}_{tk} f(y_t; \beta_k). \tag{10}$$

Accordingly, IFM proceeds by finding the parameter values $\hat{\alpha}$ and $\hat{\beta}$ that respectively maximize (9) and (10) and then maximizing function (8), evaluated at $\alpha = \hat{\alpha}$ and $\beta = \hat{\beta}$, to obtain an estimate of γ.

The procedure outlined above does not produce confidence intervals of the estimates, which however can be computed by taking a parametric bootstrap approach, by re-fitting the model from a number R of bootstrap samples, simulated from the estimated model parameters, and computing, for example, the 2.5% and the 97.5% quantiles of the empirical distribution of each bootstrap estimate.

Simulation of the model is straightforward. First, a sequence of states is simulated from a Markov chain with the desired transition probabilities, by repeatedly drawing samples from a multinomial distribution with K states. Given a sequence of states, a cylindrical observation at time t is obtained as follows. First, a sample η_1 is drawn from the uniform circular distribution. Then a sample η is drawn from the binding distribution g, evaluated at $\gamma = \hat{\gamma}_k$, where k is the state of the chain at time t. Finally, by setting $\eta_2 = (\eta + q_k \eta_1) \bmod 2\pi$, where "mod" indicates the modulus operator and q_k is the value taken by q under state k, the cylindrical sample at time t is obtained as $(x_t, y_t) = (F^{-1}(\eta_1/2\pi; \hat{\alpha}_k), F^{-1}(\eta_2/2\pi; \hat{\beta}_k))$, where k is the state of the chain at time t.

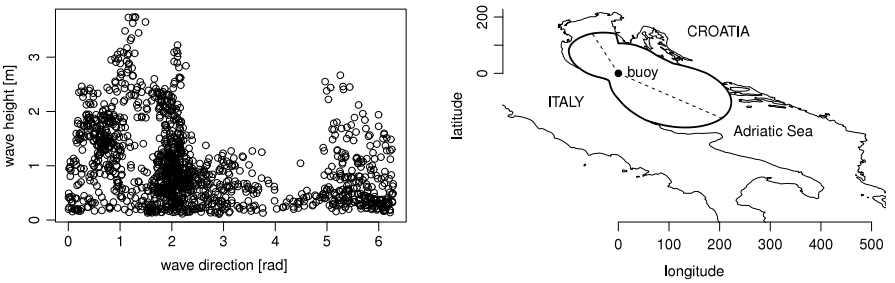

Fig. 1. Wave directions and heights, as observed by the buoy of Ancona. Left: scatterplot of wave directions and heights in winter time, as observed at the Ancona buoy; 0, $\pi/2$, π and $3/4\pi$ respectively indicate North, East, South and West. Right: the location of the buoy in the Adriatic sea; the continuous line indicates the direction-specific (smoothed) distance between the buoy and the nearest coast (known as effective fetch in marine studies); large fetch values are associated with directions where waves can move freely without orographic obstacles; the two dashed segments indicate the major axes of the Adriatic basin.

Table 1
Parameter estimates and bootstrap quantiles of a 3-state cylindrical hidden Markov model.

State 1	Parameter	Estimate	2.5% quantile	97.5% quantile
Wrapped Cauchy	Location	0.728	0.654	0.763
	Concentration	0.691	0.682	0.718
Weibull	Shape	2.267	2.198	2.301
	Scale	2.046	1.983	2.110
Copula	Dependence	0.024	0.015	0.054
	q	-1		
State 2	Parameter	Estimate	2.5% quantile	97.5% quantile
Wrapped Cauchy	Location	2.023	1.981	2.091
	Concentration	0.864	0.777	0.899
Weibull	Shape	2.033	1.971	2.145
	Scale	1.200	1.010	1.245
Copula	Dependence	0.106	0.090	0.210
	q	-1		
State 3	Parameter	Estimate	2.5% quantile	97.5% quantile
Wrapped Cauchy	Location	6.063	4.112	-2.216
	Concentration	0.137	0.110	0.156
Weibull	Shape	2.321	2.134	2.578
	Scale	0.414	0.222	0.428
Copula	Dependence	0.036	0.017	0.125
	q	1		
		Destination		
	Origin	State 1	State 2	State 3
	State 1	0.983	0.006	0.011
	State 2	0.012	0.984	0.004
	State 3	0.000	0.014	0.986

4. Application

The proposed methods have been implemented to segment a time series of semi-hourly wave directions and heights, recorded in the period 2/15/2010–3/16/2010 by the buoy of Ancona, located in the Adriatic Sea at about 30 km from the coast (Fig. 1). For simplicity, the data have been plotted on the two-dimensional Euclidean space, although they live in a cylinder. The positive direction of the angles is the clockwise direction and 0 indicates North. Segmentation of these data according to meaningful environmental regimes is often required in studies of the drift of floating objects and oil spills, in the design of off-shore structures and in studies of sediment transport and coastal erosion.

The proposed HMM requires a choice of the binding density g and a choice of the densities $f(x;\alpha)$ and $f(y;\beta)$ that respectively model the marginal distribution of the circular and the linear components of the data. The Weibull distribution is often recommended as a simple model for environmental positive variables and nicely captures the skewness of the marginal distribution of wave height. It therefore seems a natural candidate for $f(y)$. Furthermore, Jones et al. (2015) recommend to use wrapped-Cauchy specifications for both the binding density g and the marginal density $f(x)$ in order to have unimodal cylindrical densities. Unimodality of the mixture components facilitates the interpretation of the results in segmentation problems, such as the one considered here.

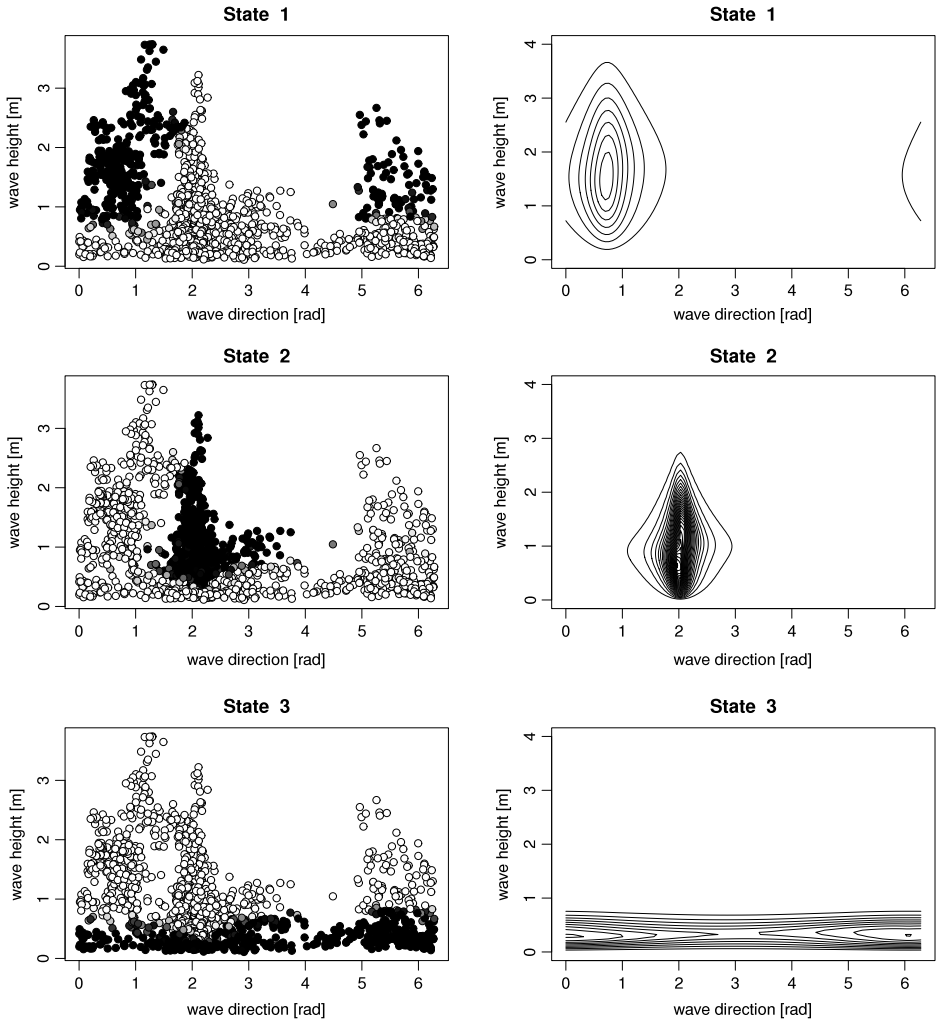

Fig. 2. Left: observations colored with gray levels according to the estimated membership probabilities of each latent state (black indicates a probability equal to 1). Right: contour plot of stare-specific copula-based densities.

By following these recommendations, the binding density has been specified as a centered wrapped Cauchy, say

$$g(v;\gamma)=\frac{1}{2\pi}\frac{1-\gamma^2}{1+\gamma^2-2\gamma\cos(v)}\quad v\in[0,2\pi).$$

This circular density depends on a single concentration parameter $\gamma\in[0,1)$ and reduces to the uniform circular density when $\gamma=0$. A wrapped Cauchy density that includes an additional location parameter α_1 has been instead exploited to model the marginal distribution of wave direction, say

$$f(x;\alpha)=\frac{1}{2\pi}\frac{1-\alpha_2^2}{1+\alpha_2^2-2\alpha_2\cos(x-\alpha_1)}\quad x\in[0,2\pi).$$

Furthermore, the origin of the cdf $F(x;\alpha)$ has been fixed at 0 (North). Finally, a Weibull distribution was exploited for wave heights, namely

$$f(y;\beta)=\frac{\beta_1}{\beta_2}\left(\frac{y}{\beta_2}\right)^{\beta_1-1}\exp\left(-\frac{y}{\beta_2}\right)^{\beta_1}\quad y\in(0,+\infty).$$

A number of models have been estimated from these data, by varying the number K of components from 2 to 4, and associating each component with $q=\pm1$. The BIC statistic suggested to segment the data according to 3 cylindrical distributions $f_{-1}(\mathbf{z};\theta_1), f_{-1}(\mathbf{z};\theta_2), f_1(\mathbf{z};\theta_3)$. Table 1 displays the estimates under these three latent classes (or states), along with bootstrap percentiles, computed by simulating 400 samples.

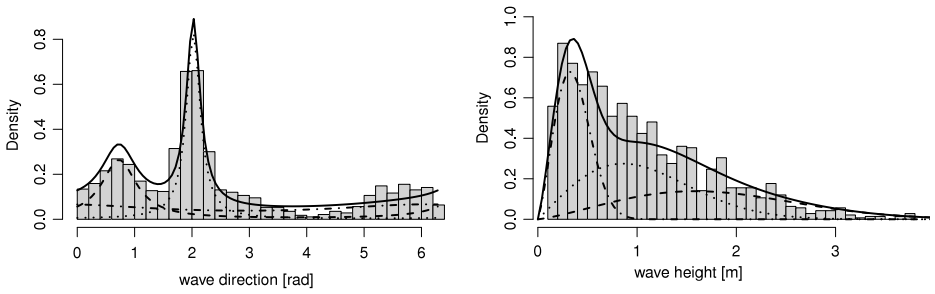

Fig. 3. Marginal distributions of the data: observed (histogram) and estimated by the model (continuous line); the dashed line, the dotted line and the dot-dashed lines indicate the estimated marginal distribution within latent state 1, 2 and 3, respectively.

The first component of the model (state 1) is associated with high waves coming from North. These waves are generated by northern Bora jets that blow along the major axis of the Adriatic basin. Under a Bora episode, most of the wind energy is transferred to the sea surface and, as a result, most of the data with the highest waves in the sample are clustered within this regime. The second component (state 2) is associated with Sirocco episodes. In this regime, waves travel southeasterly along the major axis of the Adriatic basin, driven by winds that blow from a similar direction. As expected, these two components are associated with the two major axes of the Adriatic basin where the effective fetch is large (Fig. 1; right side). The third component (state 3) is instead associated with periods of calm sea. Under this regime, moderate waves are uniformly distributed around the circle of directions.

The rows at the bottom of Table 1 include the estimated transition probabilities of the latent Markov chain. The transition probability matrix is essentially diagonal, reflecting the temporal persistence of the classes. The adopted segmentation model hence confirms that the sea surface in the study area tends to alternate relevant marine events with periods of good sea conditions.

Fig. 2 displays the contour plots of the three cylindrical densities (right side) and the scatterplot of the points, colored with gray levels that are proportional to the estimated posterior probabilities (5) of class membership (black indicates 1). Remarkably, most points are black, indicating that the model segments the data according to well separated latent classes.

By computing the proportion p_k of the data that have been MAP-allocated to each class k, we obtain the estimated marginal distributions of wave direction ($\sum_k p_k f(x, \hat{\alpha}_k)$) and wave height ($\sum_k p_k f(y, \hat{\beta}_k)$), displayed as continuous lines and respectively overlapped on their observed counterparts in Fig. 3. The reasonable goodness of fit shown by this picture can be improved by choosing a larger number K of components, if desired.

5. Discussion

A novel HMM is introduced for segmenting cylindrical times series according to a finite number of latent classes, associated with cylindrical densities that describe the distribution of the data under each class.

From a methodological viewpoint, the model offers a number of advantages. It parsimoniously accommodates temporal auto-correlation, multimodality and linear–circular correlation. It flexibly allows for any marginal distributions that is required by a specific case study. Parametric inference is relatively inexpensive from a computational viewpoint.

In the considered case study of marine data, the model segmented a time series of wave heights and directions according to intuitively appealing latent classes, providing a parsimonious description of wave dynamics in terms of interpretable environmental regimes.

Acknowledgments

F. L. was supported by the 2015 PRIN supported project ‘Environmental processes and human activities: capturing their interactions via statistical methods’, funded by the Italian Ministry of Education, University and Research.

References

Bulla, J., Lagona, F., Maruotti, A., Picone, M., 2012. A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. J. Agric. Biol. Environ. Stat. 17 (4), 544–567.

García-Portugués, E., Crujeiras, R., González-Manteiga, W., 2013. Exploring wind direction and SO2 concentration by circular-linear density estimation. Stoch. Environ. Res. Risk Assess. 27 (5), 1055–1067.

Holzmann, H., Munk, A., Suster, M., Zucchini, W., 2006. Hidden Markov models for circular and linear-circular time series. Environ. Ecol. Stat. 13, 325–347.

Johnson, R.A., Wehrly, T.E., 1978. Some angular-linear distributions and related regression models. J. Amer. Statist. Assoc. 73, 602–606.

Jones, M.C., Pewsey, A., Kato, S., 2015. On a class of circulas: copulas for circular distributions. Ann. Inst. Statist. Math. 67 (5), 843–862.

Kim, G., Silvapulle, M., Silvapulle, P., 2007. Comparison of semiparametric and parametric methods for estimating copulas. Comput. Statist. Data Anal. 51, 2836–2850.
Lagona, F., Picone, M., Maruotti, A., 2015a. A hidden Markov model for the analysis of cylindrical time series. Environmetrics 26 (8), 534–544.
Lagona, F., Picone, M., Maruotti, A., Cosoli, S., 2015b. A hidden markov approach to the analysis of space-time environmental data with linear and circular components. Stoch. Environ. Res. Risk Assess. 29 (2), 397–409.
Ley, C., Verdebout, T., 2017. Modern Directional Statistics. Chapman & Hall/CRC, London UK.
Mastrantonio, G., Maruotti, A., Jona-Lasinio, G., 2015. Bayesian hidden Markov modelling using circular-linear general projected normal distribution. Environmetrics 26 (2), 145–158.